

Methodological Aspects in Studies Based on Clinical Routine Data

Lieven Nils Kennes

Received: May 30, 2017 / Published online: September 12, 2017
© The Author(s) 2017. This article is an open access publication

ABSTRACT

Abstract: Randomized controlled clinical trials are regarded as the gold standard for comparing different clinical interventions, but generally their conduct is operationally cumbersome, time-consuming, and expensive. Studies and investigations based on clinical routine data on the contrary utilize existing data acquired under real-life conditions and are increasingly popular among practitioners. In this paper, methodological aspects of studies based on clinical routine data are discussed. Important limitations and considerations as well as unique strengths of these types of studies are indicated and exemplarily demonstrated in a recent real-case study based on clinical routine data. In addition two simulation studies reveal the impact of bias in studies based on clinical routine data on the type I error rate and false decision rate in favor of the inferior intervention. It is concluded that correctly analyzing clinical routine data yields a valuable addition to clinical research; however, as a result of a lack

of statistical foundation, internal validity, and comparability, generalizing results and inferring properties derived from clinical routine data to all patients of interest has to be considered with extreme caution.

Funding: Grünenthal GmbH.

Keywords: Bias; Blinding; Clinical routine data; Internal validity; Randomization

INTRODUCTION

In clinical research, randomized controlled clinical trials are regarded as the gold standard for comparing different clinical interventions (CONSORT statement [1]). Conducting a clinical trial and the acquisition of patient data are very cumbersome, time-consuming, and expensive. Furthermore, randomized controlled clinical trials are occasionally criticized as being too artificial and unrealistic and sometimes even for generating non-reproducible results [2, 3]. On the contrary, in clinical routine, a vast amount of data is gathered under real-life conditions on a daily basis. Detailed information on baseline characteristics, treatments, exposures, and outcomes is assessed on an individual level. Providing a platform to collect, centralize, aggregate, and store these data naturally represents a great solution to the problems involved with data acquisition in clinical trials. For the treatment of pain, for example, the German

Enhanced content To view enhanced content for this article go to <http://www.medengine.com/Redeem/CDFBF060512A56EE>.

L. N. Kennes (✉)
Department of Economics and Business
Administration, University of Applied Sciences
Stralsund, Stralsund, Germany
e-mail: lieven.kennes@hochschule-stralsund.de

Pain Practice Registry (<http://dgschmerzmedizin.de/schmerzdokumentation/praxisregister.htm>) and the related online documentation service iDocLive® (<https://info.idoclive.de/>) provide an excellent platform to centrally collect, store, and process daily routine data. These data can now be utilized and analyzed to derive valuable information.

Studies based on clinical routine data have several advantages. It is usually less difficult to realize higher sample sizes, enabling the investigation of smaller differences. Furthermore, these studies may include patients that are usually excluded from clinical trials, e.g., with a wider range of exposure levels [4]. Most of all, however, practitioners praise the time- and cost-effective approach of data acquisition as well as the strong connection to actual clinical practice.

While the time/cost argument is clearly important and always to be considered, it is merely connected to external factors and not to the actual research question at stake: discovering the true benefit of a (or several) clinical intervention(s). The argument concerning the connection to actual practice is also not directly applicable to the actual research question. Clearly, the intention and primary purpose of medical research is to help patients in practice. However, just because treatments have a specific reputation and are applied in a specific way in practice, does this mean it is correct and optimal to do so?

Unambiguously, the answer is no and it would contradict clinical development. Let us imagine, in clinical routine, that drug A is used primarily on a slightly different subset of patients to drug B. Solely comparing results of those drug interventions yields a biased picture and between-group results cannot be accredited to actual treatment difference, but rather to heterogeneity of treatment groups. Even if treatment groups are comparable at baseline, clinical routine might treat interventions differently with respect to concomitant medications, drug titrations, dose levels, or other aspects discussed later in this article. Again, results of drug intervention cannot be accredited to the drug itself but rather to those other factors.

As a result of this lack of objectivity, comparability, and internal validity, randomizing subjects to treatments is indispensable to obtain valid statistical results. Randomization is not only one of the most important techniques to avoid bias in clinical trials [5] but it also provides a basis for quantitative evaluation of clinical trial data [6]. Thus experts in the field of clinical trial design increasingly stress the importance of randomization and blinding (e.g., [7, 8]).

This article will discuss methodological issues in evaluations based on clinical routine data, pointing out strengths and limitations. This article does not contain any new studies with human or animal subjects performed by the author. The “Results” section will first elaborate general aspects of clinical routine data, followed by an illustration of these strengths and limitations in a recent real-case example and results of a simulation study. Conclusions are drawn at the end of this article.

RESULTS

General Aspects of Clinical Routine Data

Centrally storing and merging data gathered during clinical routine overcomes many difficulties in acquiring clinical data. Processing these data yields a unique opportunity to obtain a comprehensive and aggregated insight into the current state of treatment and care of a specific subset of patients:

- How many of those patients receive a specific treatment?
- What dose levels are administered in practice?
- What concomitant medications are used?
- ...

It also gives insight into the differences between daily life and clinical trials regarding how treatments are applied, what patient populations look like, etc. All this information is without doubt highly valuable. In statistical terms, all these investigations are of descriptive nature; thus, these questions can be answered with descriptive statistics. Inferential statistics (e.g., statistical hypotheses tests deriving estimates and p values) on the other hand use (a

small set of) observed data to infer properties about a larger population. Those generalized conclusions, e.g., on actual treatment differences, however, might be highly questionable if they are based on data gathered during clinical routine because of several methodological aspects. These aspects will be discussed in more detail in the remainder of this section.

Population

In an ideal world, data on every single treatment would be assessable. A truly random sample of this population would yield a sound and valid basis for statistical inference about this population [6]. However, this situation is not available in practice and thus the existing subset of clinical routine data has to be critically investigated:

- Is there a geographical bias in data collection?
- Are younger or older patients more willing to consent to data storage?
- Do participating physicians systematically prefer certain treatments?
- Which data are collected for which patients?
- ...

Speaking in statistical terms (e.g., [9]), given a treatment this subset of patients can hardly ever form a representative subset of all patients of interest. Lacking the basic principle of the population model, there is no foundation for statistical inference and without proper randomization there is not even the possibility of an invoked population model [6]. Without a basis for statistical inference, results cannot be generalized to all patients of interest.

Data Collection

In clinical trials, data are collected in a pre-specified, prospective, and homogeneous way. Clinical routine data unfortunately lack most of these attributes as a result of their biggest strength—clinical practice. Not only any dose adjustment or prescriptions of concomitant medication but in fact any behavior, interaction with and treatment of the patient, order and style of data assessment, etc. do not follow a pre-specified protocol but are done at the discretion of the physician. This results not only in

high heterogeneity in the data but also opens any door to selection bias, detection bias, attrition bias, measurement bias, and many other sources of bias (e.g., [6, 7, 10, 11]). Retrospectively sampling any subset of available routine data does not overcome these issues, as heterogeneity and bias might already be part of the data. To illustrate this with a fictive example, let us imagine two treatments:

- Treatment A: Placebo.
- Treatment B: An active pain medication proven to be slightly less efficacious, but much better tolerable than morphine.

For sufficient pain relief doctors are allowed to prescribe morphine as per required need in both cases. Now for sufficient pain relief, treatment A (placebo) is always administered with a high dose of morphine as concomitant medication, while treatment B (the active and efficacious drug) is always administered without morphine as concomitant medication. Efficacy results comparing drug A and drug B might erroneously show that treatment A (placebo) is superior to treatment B. Retrospectively taking any random subsample of patients treated this way will not prevent this false result as heterogeneity and bias are already part of every assessment and thus included in all data.

As a consequence, observed treatment difference cannot be accredited to the actual treatments but to a mixture of treatment difference, heterogeneity, and biases. An illustrative real-case example regarding systematic different use of concomitant medication in clinical routine data is given below.

Randomization

Randomization is the most important design technique in clinical investigations, providing three important benefits [12]:

- (a) Randomization is used to balance not only known but also unknown and unobservable (latent) covariates [13]. Investigators often try to “demonstrate” baseline equality of covariates, even though it is from a statistical perspective illogical and the CONSORT Group explicitly discourages this behavior [14]. In addition, testing of baseline differences is only feasible for observable covariates but is impossible for

unknown and latent covariates. As seen above, homogeneity between treatment groups is imperative for investigating true treatment difference.

- (b) Given the specific challenges in clinical research, it is mostly impossible to create a truly representative subset of patients of interest. Thus randomly assigning patients to treatment groups is imperative to provide a basis for statistical inference [6].
- (c) Furthermore in combination with blinding (see below), randomization helps to avoid bias in clinical investigations [5].

Overall, true randomization is the only technique enabling one to investigate a clinical hypothesis with internal validity. It is important to note that retrospectively taking a random selection of available routine data is not in line with ICH E9 [5], does not provide a basis for statistical inference about all patients of interest, does not create homogeneity between treatment groups, and does not avoid bias.

Blinding

Randomization helps to create homogeneity in the data and avoid bias. However, without blinding, randomization lacks the capability to prevent many types of bias [5]. Detection/ascertainment bias, attrition bias, performance bias, co-intervention bias, and observer bias [11] can simply not be prevented by randomization without blinding. Even selection bias, commonly believed to be fully avoided by randomization, is a threat without blinding [15–17]. First- and second-order selection bias [18] can be prevented by randomization with allocation concealment. However, even in the case of proper randomization, without blinding (in this case masking past treatment assignments to the investigator), third-order selection bias can lead to substantial type I error rate elevation and thus lead to false test decisions [16, 17]. Thus, even in a randomized clinical trial, without blinding, selection bias has to be regarded and, if present, corrected for appropriately [10, 19–21].

Proper blinding and preventing the forms of bias described above refers to blinding the patient and the investigator. It is important to note that conducting a blinded analysis, i.e.,

blinding only the statistician during the analysis, does not prevent any of the above biases.

Random Selection of Existing Clinical Data

As described above, randomly selecting a subset of existing clinical routine data collected in the past does not prevent bias, does not increase homogeneity in known or latent covariates, and does not provide a basis for statistical inference. Even worse, considering publication bias, in a retrospective analysis, theoretically one could successively take multiple random selections until the desired result is available and only publish results on that specific “random” sample.

In sampling theory, there are further operative reasons for taking a random sample of a larger population [9]. Taking a subsample is always accompanied by a loss of information; however, it is often not possible or at least extremely time- and cost-consuming to acquire data on all individuals of interest (e.g., election forecasts, opinion polls, etc.). Sometimes it is not even reasonable to assess all data (e.g., chemical investigation on the quality of a shipment of fruits). Those reasons all apply to the acquisition of information (data). When information (data) is already gathered and available, as for collected and stored clinical routine data, there is no operative reason to take a subsample. Analyzing a larger data set requires neither more time nor additional costs, but it does yield more information.

Use of Intention-to-Treat Principle (ITT)

There is a strict need for an ITT population for confirmatory pivotal clinical trials [5]. However, the ITT principle is unambiguously defined as including all randomized subjects in the analysis and thus inseparable from true randomization. Using the term ITT in a non-randomized investigation is inappropriate and misleading. Thus studies based on clinical routine data should abstain from referring to an ITT population. Clearly the label ITT has achieved a certain standing and importance; thus, trial design experts criticize this misuse of the ITT labeling in clinical investigations as “simply dishonest” [8].

Analyses of Clinical Routine Data in the Ueberall and Mueller-Schwefe Publication [22]

In the recent past, a study based on clinical routine data investigated the efficacy and tolerability balance of oxycodone/naloxone and tapentadol in chronic low back pain with a neuropathic component [22]. Even though the title and various parts of the Ueberall and Mueller-Schwefe [22] publication (UMS publication) repeatedly use the words “blinded”, “random”, and “prospective”, it is important to note that their investigation on clinical routine data is in fact not randomized, not blinded, and their analysis is retrospective. Consequently, all methodological issues described in Sect. 2 have to be considered carefully. Without internal validity and without a basis for statistical inference, validity of generalized conclusions is highly questionable. While the remainder of this section will be devoted to illustrate in more detail some of the aspects of Sect. 2, further explicit insight into bias in the UMS publication can be found in a response to the UMS publication by the Cochrane Group [23].

Use of Laxatives

Various studies demonstrate the favorable tolerability profile of tapentadol prolonged release (PR) compared to other opioids (e.g., [24]). Comparison of tapentadol PR (TAP) to oxycodone/naloxone PR (OXN) in a prospective randomized clinical trial treating subjects homogeneously according to a prespecified protocol proved the tolerability of TAP to be superior to that of OXN [25]. In particular, TAP is associated with significantly lower incidences of constipation than OXN. In daily clinical routine, physicians treat patients according to their individual needs. This, however, might lead to systematic differences between two treatments in the use of concomitant medication. Regarding TAP and OXN, it might be expected that in daily clinical routine, physicians systematically rely more frequently on laxatives for patients treated with OXN compared to patients treated with TAP. The subset of daily clinical routine data assessed by

iDocLive® and presented in the UMS publication actually demonstrates this behavior. Unfortunately the UMS publication erroneously claims that:

- “The proportion of patients without using laxatives changed insignificantly from baseline to the study end for both treatments” and
- “Analyses of the available patient information on the use of laxatives revealed a mixed, however, comparable, utilization pattern for both treatment groups evaluated”.

In the UMS publication, data were analyzed using very basic statistical hypotheses tests (Student’s/paired sample *t* test, χ^2 test). Unfortunately, the statistical testing procedure needed to correctly analyze the changes of laxative use, considering stochastic dependencies, is not among them. Given a baseline/end of treatment scenario, paired nominal data are present, where assessments of the same individual are stochastically dependent.

Based on information provided in the UMS publication, Tables 1 and 2 illustrate the concordant (main diagonal) and discordant (secondary diagonal) pairs of baseline/end of treatment laxative use for OXN and TAP, respectively. The discordant pairs separate perfectly in opposite directions: while the discordant pair in Table 1 (OXN) reflects perfect separation towards the additional need for laxatives at the end of treatment, the discordant pair in Table 2 (TAP) reflects perfect separation towards an improvement regarding laxative intake at the end of treatment.

Analyzing the discordant pair in Table 1 with McNemar’s test for paired nominal data yields a significant difference (worsening) for OXN in

Table 1 Laxative intake with oxycodone/naloxone PR

OXN	End of treatment		Σ
	Yes	No	
Baseline yes	30	0	30
Baseline no	10	88	98
Σ	40	88	128

Table 2 Laxative intake with tapentadol PR

TAP	End of treatment		Σ
	Yes	No	
Baseline yes	33	5	38
Baseline no	0	95	95
Σ	33	100	133

laxative intake comparing baseline to the end of treatment ($p = 0.001565/p = 0.004427$, without/with Edward's continuity correction, respectively). Analyzing the discordant pair in Table 2 shows a clear trend of improvement from baseline to the end of treatment for TAP ($p = 0.02535/p = 0.07364$, without/with Edward's continuity correction, respectively).

Striving for a simple but direct comparison of OXN to TAP, taking the development of each individual patient into account, for each patient it could be assessed whether the use of laxatives improved, worsened, or did not change through the course of treatment (Table 3). Analysis of the data in Table 3 with Fisher's exact test demonstrates a significant difference between OXN and TAP ($p < 0.0001$) in favor of TAP.

Baring in mind the general aspects of statistical analyses on clinical routine data in Sect. 2, one has to interpret these analyses as a mere description of the specific subset of daily routine data assessed by iDocLive®. However, the UMS publication chose a composite efficacy/tolerability endpoint for their inferential primary statistical analyses. This primary endpoint is heavily affected by constipation and

Table 3 Change in need for laxative in patients administered tapentadol PR and oxycodone/naloxone PR

	OXN	TAP	Σ
Improvement	0	5	5
No change	118	128	246
Worsening	10	0	10
Σ	128	133	261

laxative intake. Thus, this example demonstrates that regarding their primary endpoint, observed treatment difference cannot be accredited solely to the actual treatments. Tolerability of OXN was supported by systematic disproportionate use of concomitant medications preventing a valid comparison between OXN and TAP.

Strengths of Clinical Routine Data

The above example demonstrates that observed treatment difference can often not be accredited to the actual treatments but to a mixture of treatment difference, heterogeneity, and biases. It also illustrates the unique strength and benefit of analyzing routine data, giving a detailed insight into the current treatment and care of patients in actual clinical practice.

The limitations on the population described in Sect. 2 still have to be considered; however, data can be utilized to give a detailed picture of patient treatment by the specific non-random subset of broad-specified pain specialists. As seen above, it reveals, for instance, that in actual clinical practice for the considered subset, laxatives are used more frequently with OXN than with TAP. Furthermore, the UMS publication reveals valuable descriptive information on baseline conditions and average doses that are administered in real life as opposed to clinical trials. Most information in the publication is about the small selection of routine data (261 of 579 "appropriate" patients). As discussed in Sect. 2, looking at the whole picture achieves an even broader insight without loss of information. This particularly applies to investigations into how many patients are treated with either (or even different) pain medication(s).

Simulation Study

A simulation study was conducted to investigate the potential impact of bias in studies based on clinical routine data. Results of this simulation study show not only a substantial type I error rate elevation but also illustrate that an inferior drug might actually demonstrate superiority with a high likelihood.

Statistical Model

In a parallel group design, clinical routine data on drug A are compared to clinical routine data on drug B. The variable of interest, i.e., the response variable, is continuous and assumed to be normally distributed, with homoscedasticity between treatment groups. The true and unbiased treatment effects of drug A and drug B are denoted by μ_A and μ_B , respectively.

Clinical routine data are prone to all types of bias described in Sect. 2. Types and magnitude of bias in studies based on clinical routine data might be very different from study to study. Which types of bias are actually present in a particular study of this nature is usually difficult to detect and quantify. In statistical terms, next to the treatment effects μ_A/μ_B , the magnitude of this “bias effect” always has to be related to the standard deviation and the resulting quotient is denoted by γ . The literature is quite different when it comes to the magnitude of these bias effects. Regarding selection bias, for example, Proschan [17] uses a bias effect of $\gamma \in \{0.1, \dots, 0.5\}$, Berger et al. [15] use $\gamma \in \{0, 0.5, \dots, 2\}$, and Follmann and Proschan [26] use $\gamma = 1$. In this simulation study, the bias effect is not necessarily reflecting only a single type of bias (e.g., selection bias), but might also reflect the sum of different biases. For example, the Cochrane Group sees a recent real-case study based on clinical routine data, the UMS publication, to be “at serious risk of bias” and graded various types of bias as “moderate”, “problematic”, or of “serious concern” [23]. Nevertheless, to be conservative in this simulation study, light bias effects of $\gamma \in \{0.1, 0.3, 0.5\}$ are investigated.

Studies based on clinical routine data often rely on data gathered at various sites and/or by various investigators. As a result of certain treatment characteristics or reputations, a systematic bias applying to most patients is not necessarily unlikely. However, given a multi-center setting and regarding various types of bias, different proportions of patients affected by bias are investigated. In this simulation study, for every patient a Bernoulli-distributed random variable $B_i \sim \text{Ber}(p)$ will be used to determine whether patient i was affected by bias

or not. The distribution parameter $p \in [0, 1]$ will vary in intervals of 0.1.

Finally, let ξ_i denote the treatment group of patient i , i.e., $\xi_i = -1$ if patient i receives drug A and $\xi_i = 1$ if patient i receives drug B. As a result of the lack of randomization, ξ_i is not a random variable, but rather a mere descriptor, which might be confounded by patient characteristics as well. However, this confounding would lead to an additional bias, which will be part of and thus covered by the bias effect γ . Let n_A denote the number of patients receiving drug A and n_B denote the number of patients receiving drug B.

Combining the above information and model assumptions in a joint model, one observes that the continuous response Y_i of patient i follows a normal distribution with variance σ^2 and conditional expected value

$$E(Y_i|B_i) = \frac{1 - \xi_i}{2} \mu_A + \frac{1 + \xi_i}{2} \mu_B - \xi_i B_i \gamma \sigma.$$

The conditional density function of the response Y_i of the i th patient is thus given by

$$f_{Y_i|B_i}^{(\theta)}(y_i|b_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \times \exp\left(-\frac{1}{2\sigma^2} \left[y_i - \left(\frac{1 - \xi_i}{2} \mu_A + \frac{1 + \xi_i}{2} \mu_B - \xi_i b_i \gamma \sigma \right) \right]^2 \right), \tag{1}$$

with $\theta := (\mu_A, \mu_B, \gamma, \sigma)^T$ and b_i the realization of B_i .

Part 1: No Treatment Difference

The first part of this simulation study investigates the case of no actual treatment difference, i.e., $\mu_A = \mu_B$. Furthermore $\sigma = 1$ is chosen and, as stated above, $\gamma \in \{0.1, 0.3, 0.5\}$ and $p \in \{0, 0.1, \dots, 1\}$ is investigated. In each simulation run, a study is simulated with $n_A = n_B = 130$ patients in each treatment group and responses according to Eq. (1). For each parameter combination (γ, p) , 10,000 studies are simulated and data analyzed conducting a two-sample, two-sided t test at a significance level of $\alpha = 0.05$. Since in part 1 of this simulation study, there is actually no treatment difference, i.e., $\mu_A = \mu_B$, without bias, the number of significant study results should be approximately 5% of 10,000,

i.e., approximately 500 simulated studies should falsely show a significant difference between drug A and drug B.

Simulations were carried out in R [27] (a language and environment for statistical computing, version 3.3.1) using RStudio [28] (Integrated Development for R, version 0.99.891). Results of part 1 of this simulation study are given in Table 4 and illustrated in Fig. 1.

The results of this simulation study illustrate the severe impact of bias in a clinical investigation. If no patients are affected by bias ($p = 0$), the type I error rate is approximately 5% independent of the magnitude of the bias effect (i.e., for all three values of $\gamma \in \{0.1, 0.3, 0.5\}$). Thus the hypothesis test holds the significance level as it is supposed to. The significance level is indicated by the horizontal line in Fig. 1.

If few patients (proportion $p \leq 0.2$) are affected by a very small bias effect ($\gamma = 0.1$), the type I error rate is somewhat acceptable (even though also augmented). In all other parameter combinations, there is a substantial to severe type I error rate elevation. Especially for $\gamma \in \{0.3, 0.5\}$ the results are alarming. The statistical hypotheses test erroneously indicates a significant treatment difference most of the time, even though in reality there is none. For example, for $\gamma = 0.5$ and if 40% of patients are affected by bias, the wrong test decision in favor of drug A was made in almost 90% of the simulated studies. But even if there is only a very small ($\gamma = 0.1$) but systematic (proportion $p \geq 0.9$) bias, the type I error rate is over 30% and thus unacceptably large.

Part 2: Drug A Inferior to Drug B

The second part of this simulation study investigates the case that drug A is actually inferior to

drug B, i.e., without loss of generality $\mu_A < \mu_B$. Now the impact of bias on the test decision depends on the magnitude of the bias effect in comparison to the treatment effect size. If the true relative treatment difference is larger than the bias effect, the study will still be able to come to a correct conclusion with a likelihood naturally depending on the actual difference between treatment effect size and bias effect. However, if one drug is actually better than the other but the sum of biases surpasses the treatment difference, the test decision can lightly swing in the opposite direction. The second part of the simulation study uses the same model assumptions as the first part, with the slight modification that drug A is actually inferior to drug B, i.e., $\frac{\mu_B - \mu_A}{\sigma} = 0.3$ and the bias effect surpasses this effect size, i.e., $\gamma = 0.5$. Instead of a two-sided hypothesis test at the 5% significance level, a one-sided, two-sample t test is conducted at the 2.5% significance level. The null hypothesis states that drug A is inferior or equal to drug B, and the alternative hypothesizes that drug A is superior to drug B (the latter actually being incorrect).

Results of this part of the simulation study again demonstrate that bias can have a severe impact on the test decision. Even though drug A is actually inferior to drug B, a systematic bias ($p \geq 0.7$) can easily revert this trend in the opposite direction. The proportion of wrong test decisions in favor of (the inferior) drug A is above 80% (Table 5).

CONCLUSIONS

Retrospective as well as prospective clinical studies can be influenced by various types of heterogeneity and bias. As extensively discussed

Table 4 Relative number of significant study results in the case of no treatment difference (type I error rate)

	$p = 0$	$p = 0.1$	$p = 0.2$	$p = 0.3$	$p = 0.4$	$p = 0.5$	$p = 0.6$	$p = 0.7$	$p = 0.8$	$p = 0.9$	$p = 1$
$\gamma = 0.1$	0.0501	0.0516	0.0630	0.0775	0.0960	0.1212	0.1529	0.2008	0.2525	0.3026	0.3617
$\gamma = 0.3$	0.0484	0.0750	0.1576	0.3005	0.4752	0.6619	0.8151	0.9161	0.9707	0.9920	0.9979
$\gamma = 0.5$	0.0490	0.1243	0.3524	0.6490	0.8830	0.9721	0.9961	1	1	1	1

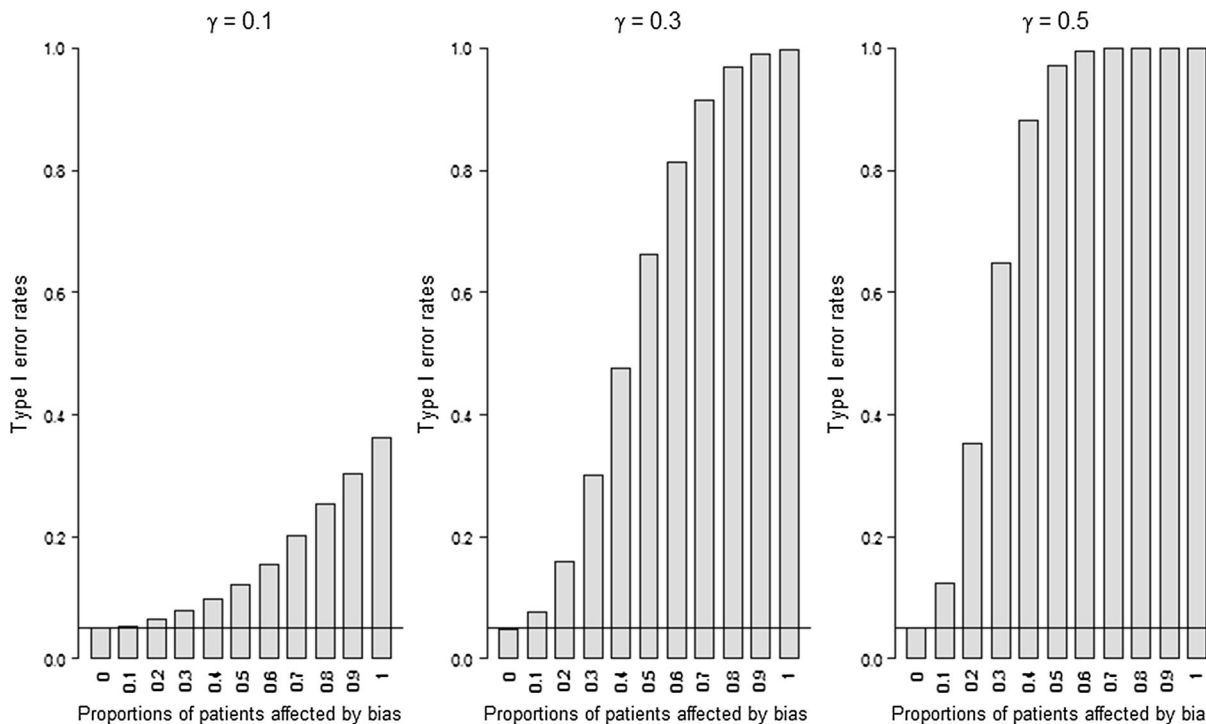


Fig. 1 Relative number of significant study results in the case of no treatment difference (type I error rate)

Table 5 Relative number of significant study results: wrong decision in favor of drug A even though drug B is superior

	$p = 0$	$p = 0.1$	$p = 0.2$	$p = 0.3$	$p = 0.4$	$p = 0.5$	$p = 0.6$	$p = 0.7$	$p = 0.8$	$p = 0.9$	$p = 1$
$\gamma = 0.5$	0.0000	0.0001	0.0015	0.0100	0.0700	0.2489	0.5278	0.8101	0.9543	0.9932	0.9994

in the literature and guidelines as well as demonstrated in the simulation study above, bias can have a severe impact on study results. Between-group results cannot be accredited to actual treatment difference, but rather to a mixture of treatment difference, heterogeneity, and bias. The impact of the last two might then not only lead to substantial impact of type I error rate elevation but is also actually able to demonstrate superiority of an inferior treatment.

In a clinical trial, multiple means and instruments exist to prevent bias and heterogeneity. Most important are randomization and blinding [5]. Thus, the CONSORT Group rightly regards randomized controlled clinical trials as the gold standard for evaluating health-care interventions (CONSORT statement [1]). Clearly not every prospective clinical trial is

randomized and blinded. Randomization might not always be possible for ethical or practical reasons, e.g., randomly assigning subjects to smoking or not smoking cigarettes. Other clinical trials, e.g., those comparing two different types of surgery, can be randomized but not double-blinded. Many prospective clinical trials have, however, at least the potential to be randomized and double-blinded. On the other hand, studies based on clinical routine data are per se not randomized and not blinded as it would contradict clinical practice. Furthermore, their analyses are usually retrospective. Taking a random selection of existing data and conducting a blinded analysis do not prevent any of the above discussed issues. Quite the contrary, taking a random selection of already existing data actually has unnecessary disadvantages compared to analyzing all data.

Without randomization, without internal validity, and without representing a random sample of the larger population of interest, a basis for statistical inference is not given. Thus there is usually no statistical foundation to generalize results and properties derived from clinical routine data to all patients of interest.

Unfortunately, the above arguments and basic concepts of probability theory are occasionally ignored by practitioners and non-statisticians, who are captivated by p values. This malpractice draws away attention of the great benefits clinical routine data actually have to offer. Descriptively summarizing clinical routine data yields a comprehensive and aggregated insight into the current state of treatment and care of a specific subset of patients. Hypotheses can be derived from this valuable insight, which can subsequently be investigated under circumstances enabling a valid comparison. Thus, like other observational studies, well-executed investigations based on clinical routine data are indispensable precursors of randomized controlled clinical trials.

In summary, prespecified, homogeneous, prospective, randomized, and blinded data acquisition is desired to obtain valid clinical trial results. By default, the acquisition of clinical routine data is usually lacking most of these attributes. However, correctly analyzing clinical routine data yields a valuable insight into the current state of treatment and care of a specific subset of patients. Fulfilling a merely descriptive purpose for a specific subset of patients, descriptive summary statistics are the method of choice. Confirmatory analyses using inferential statistics trying to generalize results to all patients of interest have to be considered carefully and are often inappropriate.

ACKNOWLEDGEMENTS

This paper was funded by Grünenthal GmbH. Article processing charges were funded by Grünenthal GmbH. The author had full access to all of the simulation study data and takes complete responsibility for the integrity of the data and accuracy of the data analysis. Data from the

UMS publication is not available to the scientific community and the author used summary information published in the UMS publication to analyze the development of laxative use of patients treated with OXN and TAP.

This manuscript was solely written by the author; no medical writing service was provided. The author meets the International Committee of Medical Journal Editors (ICMJE) criteria for authorship for this manuscript, takes responsibility for the integrity of the work as a whole, and has given final approval for the version to be published.

The author likes to thank the anonymous referees for their valuable comments and suggestions, which led to an improvement of this article.

Disclosures. Lieven Nils Kennes is professor of statistics and econometrics at Stralsund University of Applied Sciences. He holds a doctorate degree in mathematics and dedicates his research to clinical trial designs, in particular randomization and selection bias. Lieven Nils Kennes is a former employee of Grünenthal GmbH (until the beginning of 2016) and is now a consultant for Grünenthal.

Compliance with Ethics Guidelines. This article does not contain any new studies with human or animal subjects performed by the author.

Open Access. This article is distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits any noncommercial use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

REFERENCES

1. Schulz KF, Altman DG, Moher D, for the CONSORT Group. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. *BMJ*. 2010;340:c332.

2. Zwarenstein M, Oxman A. Why are so few randomized trials useful, and what can we do about it? *J Clin Epidemiol*. 2006;59:1125–6.
3. Zuidgeest M, Goetz I, Groenwold R, Irving E, van Thiel G, Grobbee DE. Pragmatic trials and real-world evidence—an introduction to the series. *J Clin Epidemiol*. 2017. doi:10.1016/j.jclinepi.2016.12.023.
4. dos Santos Silva I. *Cancer epidemiology: principles and methods*. Lyon: International Agency for Research on Cancer; 1999.
5. ICH E9. *Statistical principles for clinical trials*. 1998. Current version dated 5 Feb 1998. <http://www.ich.org>. Accessed Sep 2017.
6. Rosenberger WF, Lachin JM. *Randomization in clinical trials: theory and practice*. New York: Wiley; 2016.
7. Kahan BC, Rehal S, Cro S. Risk of selection bias in randomised trials. *Trials*. 2015;16:405.
8. Berger VW. Conflicts of interest, selective inertia, and research malpractice in randomized clinical trials: an unholy trinity. *Sci Eng Ethics*. 2015;21(4):857–74.
9. Kauermann G, Küchenhoff H. *Stichprobentheorie [sampling theory]*. Heidelberg: Springer; 2011.
10. Berger VW. The reverse propensity score to detect selection bias and correct for baseline imbalances. *Stat Med*. 2005;24:2777–87.
11. Jadad AR, Enkin MW. *Randomized controlled trials: questions, answers and musings*. New York: Wiley; 2007.
12. Armitage P. The role of randomization in clinical trials. *Stat Med*. 1982;1(4):345–52.
13. Beller EM, GebSKI V, Keech AC. Randomization in clinical trials. *Med J Aust*. 2002;177(100):565–7.
14. Moher D, Hopewell S, Schulz KF, et al. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. *J Clin Epidemiol*. 2010;63:e1–37.
15. Berger VW, Ivanova A, Knoll M. Minimizing predictability while retaining balance through the use of less restrictive randomization procedures. *Stat Med*. 2003;42:3017–28.
16. Kennes LN, Cramer E, Hilgers RE, Heussen N. The impact of selection bias on test decisions in randomized clinical trials. *Stat Med*. 2011;30:2573–81.
17. Proschan M. Influence of selection bias on type I error rate under random permuted block designs. *Stat Sin*. 1994;4:219–31.
18. Berger VW. Quantifying the magnitude of baseline covariate imbalances resulting from selection bias in randomized clinical trials. *Biom J*. 2005;47:119–27.
19. Berger VW, Exner DV. Detecting selection bias in randomized clinical trials. *Control Clin Trials*. 1999;20:319–27.
20. Ivanova A, Barrier RC, Berger VW. Adjusting for observable selection bias in block randomized trials. *Stat Med*. 2005;24:1537–46.
21. Kennes LN, Rosenberger WF, Hilgers RD. Inference for blocked randomization under a selection bias model. *Biometrics*. 2015;71:979–84.
22. Ueberall MA, Mueller-Schwefe GHH. Efficacy and tolerability balance of oxycodone/naloxone and tapentadol in chronic low back pain with a neuropathic component: a blinded end point analysis of randomly selected routine data from 12-week prospective open-label observations. *J Pain Res*. 2016;9:1001–20.
23. Duarte GS, Santos J, Costa J. Response to the publication by Ueberall and Mueller-Schwefe. *J Pain Res*. 2017;10:1055–8.
24. Lange B, Kuperwasser B, Okamoto A, et al. Efficacy and safety of tapentadol prolonged release for chronic osteoarthritis pain and low back pain. *Adv Ther*. 2010;27:381–99.
25. Baron R, Jansen JP, Binder A, et al. Tolerability, safety, and quality of life with tapentadol prolonged release (PR) compared with oxycodone/naloxone PR in patients with severe chronic low back pain with a neuropathic component: a randomized, controlled, open-label, phase 3b/4 trial. *Pain Pract*. 2015;16:600–19.
26. Follmann D, Proschan M. The effect of estimation and biasing strategies on selection bias in clinical trials with permuted blocks. *J Stat Plan Inference*. 1994;39:1–17.
27. R Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. 2016. <http://www.R-project.org>. Accessed Sep 2017.
28. R Studio Team. *RStudio: integrated development environment for R*. RStudio, Inc. 2015. <http://www.rstudio.com>. Accessed Sep 2017.